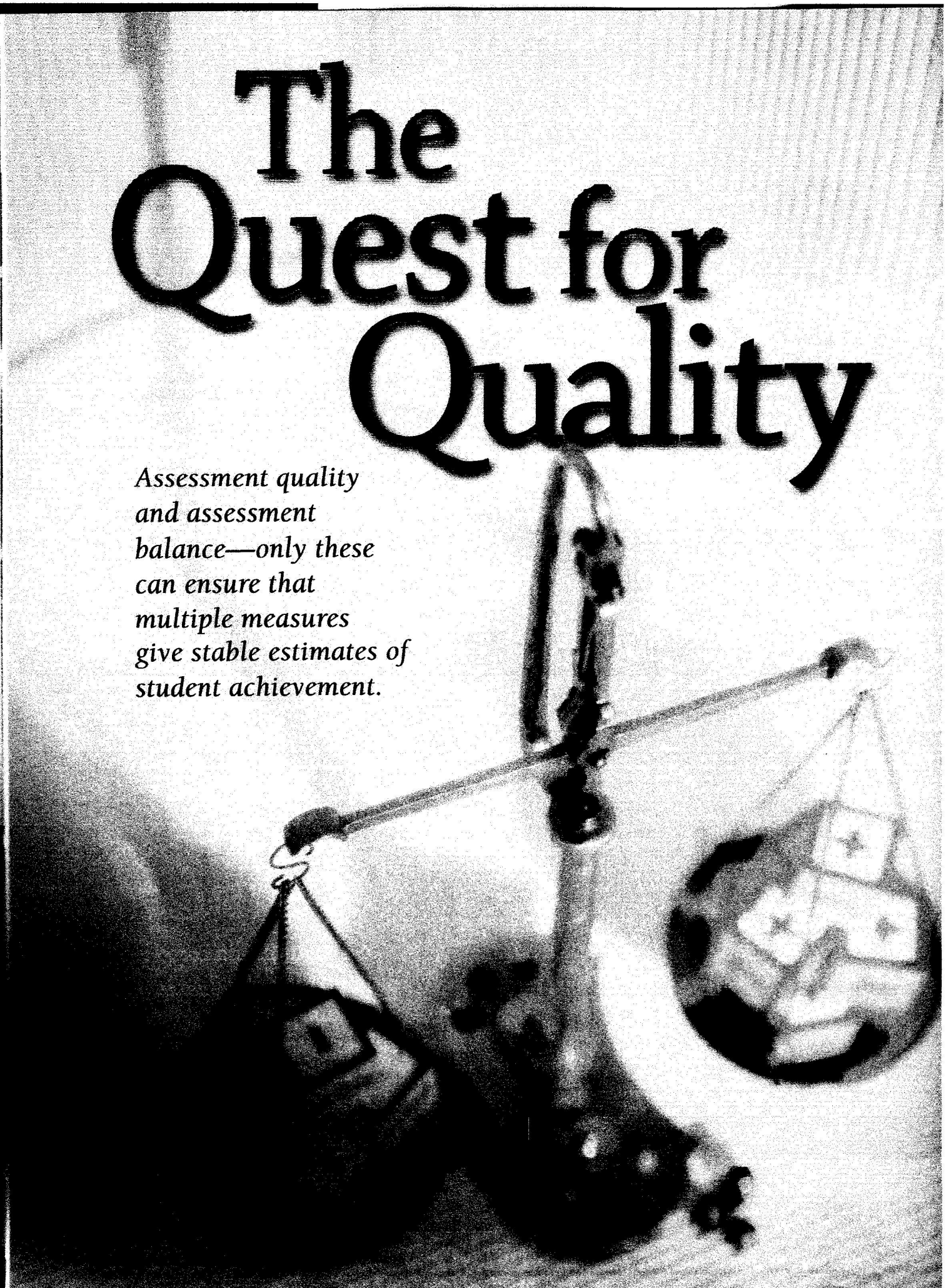
The Quest for Quality

Assessment quality and assessment balance—only these can ensure that multiple measures give stable estimates of student achievement.



Stephen Chappuis, Jan Chappuis, and Rick Stiggins

Long before No Child Left Behind (NCLB), high-stakes tests were common in schools. Cut scores on tests have dictated promotion from one grade level to the next, and teachers have used them to assign passing or failing grades. High school students continue to take course placement exams, subject-area finals, exit exams, and college entrance tests. Making decisions that affect individuals and groups of students on the basis of a single measure is part of our past and current practice.

In the past, few educators, policymakers, or parents would have considered questioning the accuracy of these tests. Most assumed that a low score or grade was probably justly assigned and that a decision made about a student as a result was as defensible as the evidence on which it was based.

But NCLB has exposed students to an unprecedented overflow of testing. In response to the accountability movement, schools have added new levels of testing that include benchmark, interim, and common assessments. Using data from these assessments, schools now make decisions about individual students, groups of students, instructional programs, resource allocation, and more. We're betting that the instructional hours sacrificed to testing will return dividends in the form of better instructional decisions and improved high-stakes test scores.

Given the rise in testing, especially in light of a heightened focus on using multiple measures, it's increasingly important to address two essential components of reliable assessments: quality and balance.

Keys to Quality

Although it may seem as though having more assessments will mean we are more accurately estimating student achievement, the use of multiple measures does not, by itself, translate into high-quality evidence. Using misinformation to triangulate on student needs defeats the purpose of bringing in more results to inform our decisions.

Five keys to assessment quality provide the larger picture into which our multiple measures must fit (Stiggins, Arter, Chappuis, & Chappuis, 2006). Only assessments that satisfy these standards—whether teachers' classroom assessments, department or grade-level common assessments, or benchmark or interim tests—will be capable of informing sound decisions.

Clear Purpose

The assessor must begin with a clear picture of why he or she is conducting the assessment. Who will use the results to inform what decisions? The assessor might use the assessment formatively—as practice or to inform students about their own progress—or summatively—to feed results into the grade book.

The use of multiple measures does not, by itself, translate into high-quality evidence.

In the case of summative tests, the reason for assessing is to document individual or group achievement or mastery of standards and measure achievement status at a point in time. The purpose is to inform others—policymakers, program planners, supervisors, teachers, parents, and the students themselves—about the overall level of students' performance.

Clear Learning Targets

The assessor needs to have a clear picture of what achievement he or she intends to measure. If we don't begin with clear statements of the intended learning—clear and understandable to everyone, including students—we won't end up with sound assessments.

For this key to quality, it's important to know the learning targets represented in the written curriculum. The four categories of learning targets are

- *Knowledge targets*, which are the facts and concepts we want students to know. In math, a

ROWITZ/GETTY IMAGES

knowledge target might be to recognize and describe patterns.

- *Reasoning targets*, which require students to use their knowledge to reason and problem solve. A reasoning target in math might be to use statistical methods to describe, analyze, and evaluate data.
- *Performance skill targets*, which ask students to use knowledge to perform or demonstrate a specific skill, such as reading aloud with fluency.
- *Product targets*, which specify that students will create something, such as a personal health-related fitness plan.

For each assessment, regardless of purpose, the assessor should organize the learning targets represented in the assessment into a written test plan that matches the learning targets represented in the curriculum.

For example, Figure 1 shows a 3rd grade math test plan. It defines what the test will cover, including such specific learning targets as being able to multiply by two (one of the learning targets in the curriculum). Creating a plan like this for each assessment helps assessors sync what they taught with what they're assessing. It also helps them assign the appropriate balance of points in relation to the importance of each target as well as the number of items for each assessed target.

Sound Assessment Design

This key ensures that the assessor has translated the learning targets into assessments that will yield accurate results. It calls attention to the proper assessment method and to the importance of minimizing any bias that might distort estimates of student learning.

Teachers have choices in the assessment methods they use, including selected-response formats, extended written response, performance assessment, and personal communication. Selecting an assessment method that is incapable of reflecting the intended learning will compromise the accuracy of the results. For example, if the teacher wants to assess knowledge mastery of a certain item, both selected-response and extended written response methods are good matches, whereas performance assessment or personal communication may be less effective and too time-consuming. Figure 2 (page 18) clarifies which assessment methods are most likely to produce accurate results for different learning targets.

Bias can also creep into assessments and erode accurate results. Examples of bias include poorly printed test forms, noise distractions, vague directions, and cultural insensitivity. Teachers can minimize bias in a number of ways. For example, to ensure accuracy in selected-response assessment formats, they should keep wording simple and focused, aim for the lowest possible reading level, avoid providing clues or making the correct answer obvious, and highlight crucial words (for instance, *most, least, except, not*).

Effective Communication of Results

The assessor must plan to manage information from the assessment appropriately and report it in ways that will meet the needs of the intended users, keeping in mind the following: Are results communicated in time to inform the intended decisions? Will the users of the results understand them and see the connection to learning? Do the results provide clear direction for what to do next?

If we don't begin with clear statements of the intended learning, we won't end up with sound assessments.

This key relates directly back to the purpose of the assessment. For instance, if students will be the users of the results because the assessment is formative, then teachers must provide the results in a way that helps students move forward. Specific, descriptive feedback linked to the targets of instruction and arising from the assessment items or rubrics communicates to students in ways that enable them to immediately take action, thereby promoting further learning.

For example, let's say the content standard you're teaching to is "Understands how to plan and conduct scientific investigations," and your assessment rubric states that a strong hypothesis includes a prediction with a cause-and-effect reason. Feedback to students can use the language of the rubric: "What you have written is a

hypothesis because it is a prediction about what will happen. You can improve it by explaining *why* you think that will happen." Or, you can highlight the phrases on the rubric that describe the hypothesis's strengths and areas for improvement and return the rubric with the work.

A grade of *D+*, on the other hand, may be sufficient to inform a decision about a student's athletic eligibility, but it is not capable of informing the student about the next steps in learning.

Selecting an assessment method that is incapable of reflecting the intended learning will compromise the accuracy of the results.

Student Involvement in the Assessment Process

Students learn best when they monitor and take responsibility for their own learning. This means that teachers need to write learning targets in terms that students will understand.

For example, suppose we are preparing to teach 7th graders how to make inferences. After defining *inference* as "a conclusion drawn from the information available," we might put the learning target in student-friendly language: "I can make good inferences. This means I can use information from what I read to draw a reasonable conclusion." If we were working with 2nd graders, the student-friendly language might look like this: "I can make good inferences. This means I can make a guess that is based on clues."

Teachers should design the assessment so students can use the results to self-assess and set goals. A mechanism should be in place for students to track their own progress on learning targets and communicate their status to others. For example, a student might assess how strong his or her thesis statement is by using phrases from a rubric, such as "Focuses on one specific aspect of the subject"or "Makes an assertion that can be argued."

Keys to Balance

The goal of a balanced assessment system is to ensure that all assessment users have access to the data they want when they need it, which in turn directly serves the effective use of multiple measures.

One way to think about the various uses of assessment in a balanced system is by grouping the assessments into levels associated with the frequency of their administration. *Ongoing classroom assessments* serve both formative and summative purposes and meet students' as well as teachers' information needs. *Periodic interim/benchmark assessments* can also serve program evaluation purposes, as well as inform instructional improvement and identify struggling students and the areas in which they struggle. *Annual state and local district standardized tests* serve annual accountability purposes, provide comparable data, and serve functions related to student placement and selection, guidance, progress monitoring, and program evaluation.

Effectively planning for the use of multiple measures means providing assessment balance throughout these three levels, meeting student, teacher, and district information needs. This is done using both formative and summative assessments, large-group and individual testing, assessing a range of relevant learning targets using a range of appropriate assessment methods.

As a "big picture" beginning point in planning for the use of multiple measures, assessors need to consider each assessment level in light of four key questions, along with their formative and summative applications[1]:

What decisions will the assessment inform?

At the level of ongoing classroom assessments, formative applications involve

FIGURE 1. A Sample 3rd Grade Math Test Plan

Learning Target	Item #s	Points
Number Sense: Identify place value to the thousands place	1, 5, 9, 11	10
Data Analysis and Probability: Interpret bar graphs	2, 6, 8, 14	12
Number Operations: Multiply by 2	3, 4, 12, 16	4
Number Operations: Subtract with regrouping	7, 10, 13, 15	4
Total Points:		**30**

what students have mastered and what they still need to learn. At the level of periodic interim/benchmark assessments, they involve which standards students are not mastering and where teachers can improve instruction right away. At the level of annual state/district standardized assessments, they involve where and how teachers can improve instruction—next year.

Summative applications refer to grades students receive (classroom level); whether the program of instruction has delivered as promised and whether the school should continue to use it (periodic assessment level); and how many students have met standards (annual testing level).

Who is the decision maker?

This will vary. The decision makers might be students and teachers at the classroom level; instructional leaders, learning teams, and teachers at the periodic level; or curriculum and instructional leaders and school and community leaders at the annual testing level.

What information do the decision makers need?

From a formative point of view, decision makers at the classroom assessment level need evidence of where students are on the learning continuum toward each standard, whereas decision makers at the next two levels want to know

FIGURE 2. Choosing the Right Assessment

LEARNING TARGET	ASSESSMENT METHOD			
	Selected Response	**Extended Written Response**	**Performance Assessment**	**Personal Communication**
Knowledge Mastery	Good match for assessing mastery of elements of knowledge.	Good match for tapping understanding of relationships among elements of knowledge.	Not a good match—too time-consuming to cover everything.	Can be used if assessor asks questions, evaluates answers, and infers mastery—but a time-consuming option.
Reasoning Proficiency	Good match only for assessing understanding of some patterns of reasoning out of context.	Written descriptions of complex problem solutions can provide a window into reasoning proficiency.	Assessor can watch students solve some problems and infer their reasoning proficiency.	Can be used if assessor asks student to "think aloud" or asks follow-up questions to probe reasoning.
Skills	Not a good match. Can assess mastery of the knowledge the students need to perform the skill well, but cannot measure the skill itself.		Good match. Assessor can observe and evaluate skills as they are being performed.	Strong match when skill is oral communication proficiency; not a good match otherwise.
Ability to Create Products	Not a good match. Can assess mastery of the knowledge students need to create quality products, but cannot assess the quality of products themselves.	Strong match only when the product is written. Not a good match when the product is not written.	Good match. Can assess the attributes of the product itself.	Not a good match.

Source: Adapted from *Classroom Assessment for Student Learning: Doing It Right—Using It Well* By R. Stiggins, J. Arter, J. Chappuis, and S. Chappuis, 2006, Portland, OR: ETS Assessment Training Institute. Copyright © 2006 by ETS. Reprinted with permission.

which standards students are struggling to master.

From a summative point of view, users at the classroom and periodic assessment levels want evidence of mastery of particular standards; at the annual testing level, decision makers want the percentage of students meeting each standard.

What are the essential assessment conditions?

These conditions are most articulated at the classroom assessment level, through the use of clear curriculum maps for each standard, accurate assessment results, effective feedback, and results that point student and teacher clearly to next steps. Summative applications at this level include accurate summaries of evidence and grading symbols that carry clear and consistent meaning for all.

At the periodic level of assessment, essential assessment conditions include results that show mastery of program standards aggregated over students. At the annual testing level, accurate evidence of how *each* student did in mastering *each* standard aggregated over students is needed.

What Assessments Can—and Cannot—Tell Us

In such an intentionally designed and comprehensive system, a wealth of data emerges. Inherent in its design is the need for all assessors and users of assessment results to be assessment literate—to know what constitutes appropriate and inappropriate uses of assessment results—thereby reducing the risk of applying data to decisions for which they aren't suited.

For example, because they understand what is appropriate at each of the three levels of assessment—both formatively and summatively—assessment-literate teachers would not

- Use a reading score from a state accountability test as a diagnostic instrument for reading group placement.
- Use SAT scores to determine instructional effectiveness.
- Rely solely on performance assessments to test factual knowledge and recall.
- Assess learning targets requiring the "doing" of science with a multiple-choice test.

Assessment literacy is the foundation for a system that can take advantage of a wider use of multiple measures. At the classroom level, teachers can choose among the four assessment methods (selected-response, extended written response, performance assessment, and personal communication). Most assessments developed beyond the classroom rely largely on selected-response or short-answer formats and are not designed to meet the daily, ongoing information needs of teachers and students. As such, not only are they limited in key formative uses, but they also cannot measure more complex learning targets at the heart of instruction.

Because classroom teachers can effectively use all available assessment methods, including the more labor-intensive methods of performance assessment and personal communication, they can provide information about student progress not typically available from student information systems or standardized test results. The classroom is also a practical location to give students multiple opportunities to demonstrate what they know and can do, adding to the accuracy of the information available from that level of assessment.

A Solid Foundation for a Balanced System

Educators are more likely to attend to issues of quality and serve the best interests of students when we build balanced systems, with assessment-literate users. From that foundation we can develop coordinated plans for the use of multiple measures, taking advantage of dependable data generated at every level of assessment. EL

Assessment literacy is the foundation for a system that can take advantage of a wider use of multiple measures.

[1]A detailed chart listing key issues and their formative and summative applications at each of the three assessment levels is available at www.ascd.org/ASCD/pdf/journals/ed_lead/el200911_chappius_table.pdf

References

Chappuis, J. (2009). *Seven strategies of assessment for learning.* Portland, OR: Educational Testing Service.

Stiggins, R., Arter, J., Chappuis, J., & Chappuis, S. (2006). *Classroom assessment for student learning—Doing it right, using it well.* Portland, OR: Educational Testing Service.

Stephen Chappuis (schappuis@ets.org), **Jan Chappuis** (jchappuis@ets.org), and **Rick Stiggins** (rstiggins@ets.org) work with the ETS Assessment Training Institute in Portland, Oregon (www.ets.org/ati).